IE6600 - hw2

Due on 10/19/2021 11:59pm ET

Zhenyuan Lu

IE6600 Homework Instructions

You should process your homework in R Markdown `(.Rmd)`, and **knit** it to `.pdf` file (do not include any data or other materials). Attentively check all the references I mentioned in the class or from other resources.

Once the homework is completed, you need to have your homework compressed into one `.zip` file (hw2YourFullName.zip), and submit it to the assignment section on Canvas. In the `.zip` file, it should contain the following documents:

hw2YourFullName.rmd

hw2YourFullName.pdf

Please include all your codes and results for each of the problem, and keep them organized and clear. All of your codes should run successfully. Problems related to any plots/charts should be generated by `ggplot2` mainly. If it's neccessary, please deal with the missing values, overplotting, or labels on axis/legend properly. **All the solutions may vary.**

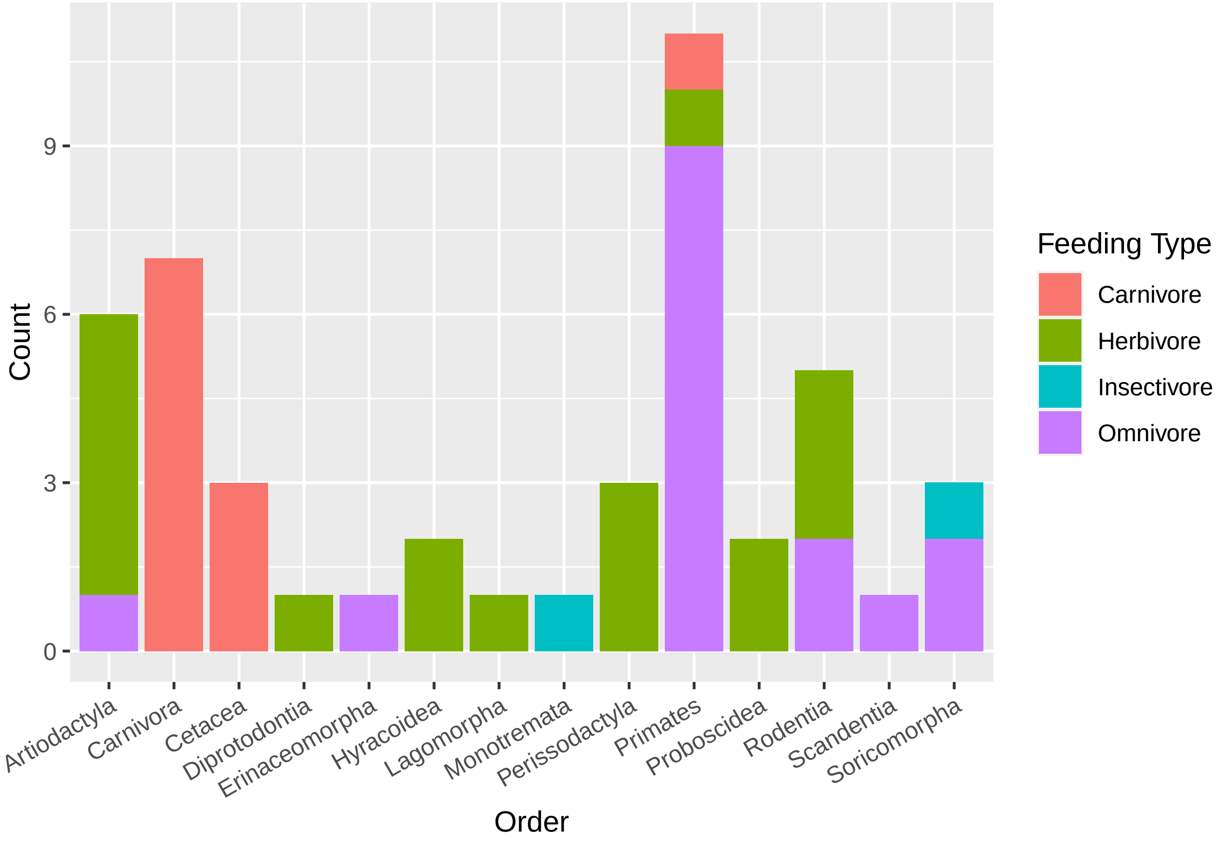
Section A

Only use ggplot2 for plotting This section is for testing your ggplot2, and data exploration skills. Dataset `msleep` from ggplot2 package will be using through this section. Use ? to check the documentation of `msleep`.

Problem 1

We are interested in those animals whose **awake time over 12 hours**. Create a bar chart as the following figure. Remove the NA values from **feeding types**: carnivore, omnivore, insectivore and herbivore.

hints: You may adjust the angel of x-axis label by using *theme(axis.text.x=element_text())*, and the legend labels by using *scale_fill_discrete()*.



Problem 2

We would like to investigate how's the relationship between total amount of sleep (hr) and brain weight(kg) among **feeding types**: carnivore, omnivore, insectivore and herbivore. Plot **total amount of sleep (hr)** versus **brain weight (kg)**, applying color mapping on the **feeding types**(vore). Remove the NA group from **feeding types**. Include a smoothing line on the plot. What do you notice in the plot?

Please deal with labels on axis/legend properly.

Problem 3

Still working on the above plot. Apply log transformation on the brain weight **Brain Weight (Kg), Log**, what do you observe in the plot?

Section B

Only use ggplot2 for plotting

Section B uses FY 2019 H-1B Employer Data from U.S. Citizenship and Immigration Services. Download FY2019 H-1B data from: https://www.uscis.gov/tools/reports-studies/h-1b-employer-data-hub-files

To read the data manual: https://www.uscis.gov/tools/reports-studies/understanding-our-h-1b-employer-data-hub

The H-1B is a visa in the United States under the Immigration and Nationality Act, section 101(a)(15)(H) that allows U.S. employers to temporarily employ foreign workers in specialty occupations. A specialty occupation requires the application of specialized knowledge and a bachelor's degree or the equivalent of work experience.

Use read.csv() to import the dataset to R.

Problem 1

Import the H-1B data.

- You may notice the data types of "Initial.Approvals", "Initial.Denials", "Continuing.Approvals", and "Continuing.Denials" are wrong. We need to convert them into numerical columns.
- Return a data frame containing the top 5 employers which have the most cases of initial approved H-1B. This data frame should have the columns: employer, initial approvals, initial denials, continuing approvals, and continuing denials. Show the top 5 data frame.
- Plot a bar chart of Employer versus Initial approvals, mapping Initial Denials as fill, what do you notice based on the plot?

Hint1: All the variables should be be associated with the proper data types

Hint2: Using function gsub() to eliminating the "," for every three decimal places. e.g. 1,000 to 1000

Hint3: When converting data from factor to numeric, be aware of the values

Problem 2

Download geocode data: https://northeastern.instructure.com/courses/91043/files/11702036/download?download_frd=1

If this link doesn't work, please go to Canvas - Home - Homework - usZipGeo.csv

- Join H-1B data table with geocode data table by State and Zip columns.
- This new data frame should include columns: zip, employer, initial approvals, initial denials, continuing approvals, continuing denials, state, city, longitude, and latitude.
- Insert a new column **prop** into this new data frame by the formula: inital denial/initial approval

Hint1: When joining two tables, make sure all the key column names are the same from both tables.

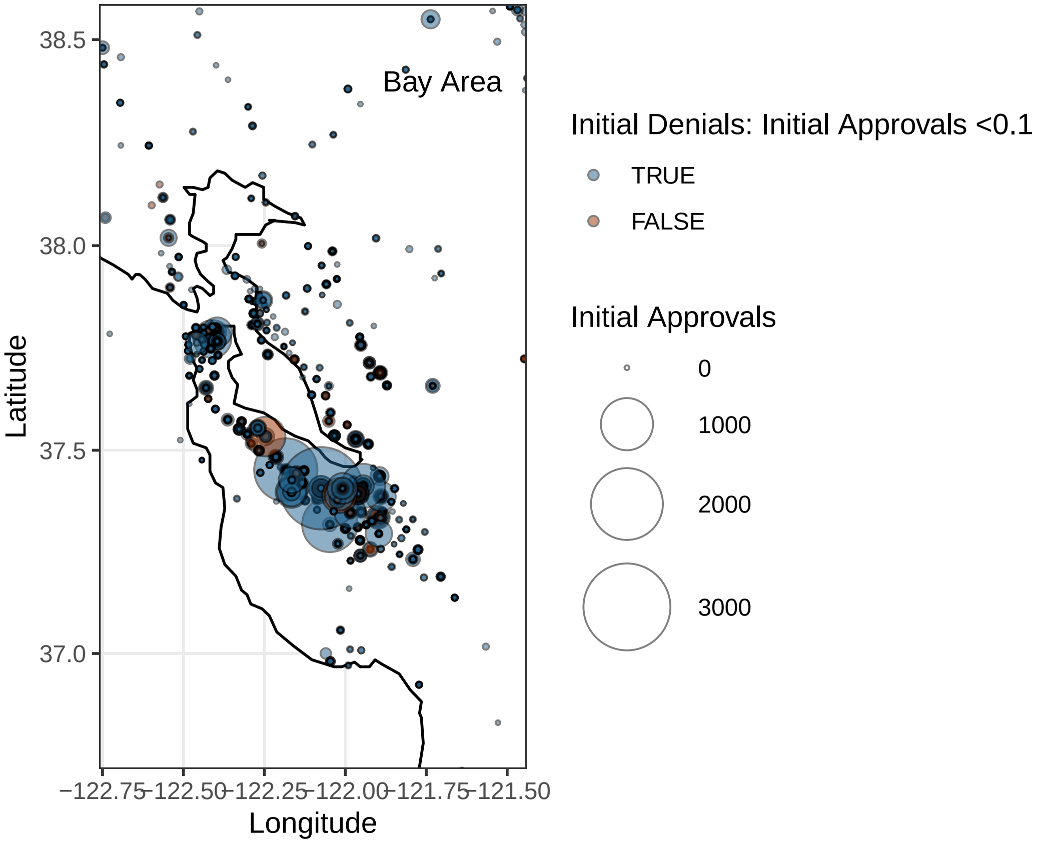
Problem 3

We are interested in the H-1B cases around Bay Area, California.

Create a map of the California, and then adjust the plotting x/y limits to a proper zoom level of Bay Area. Then showing the locations of each employer along with, the **prop** less than 0.1 (mapped as the color/fill), and the **initial approvals** (mapped as the size).

hints: Install **maps** and **mapproj** packages, and use the ggplot2::map_data() to draw "California" region of the US. Using coord_map() to set up the view range of your map.

Answer may vary
The following example is just for your reference. If you plot is slightly different from it, you should be fine.



Bay Area
Latitude
38.5
38.0
37.5
37.0
Longitude
−122.75
−122.50
−122.25
−122.00
−121.75
−121.50
Initial Denials: Initial Approvals <0.1
TRUE
FALSE
Initial Approvals
0
1000
2000
3000